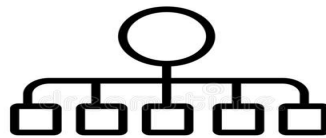

Classification On Haberman Cancer Survival Dataset.



classification



Submitted To

Prof. Uma Shanker Tiwary

Submitted By

Anand Geed

IDS2022005

(M. Tech - DSA)

Table Of Contents

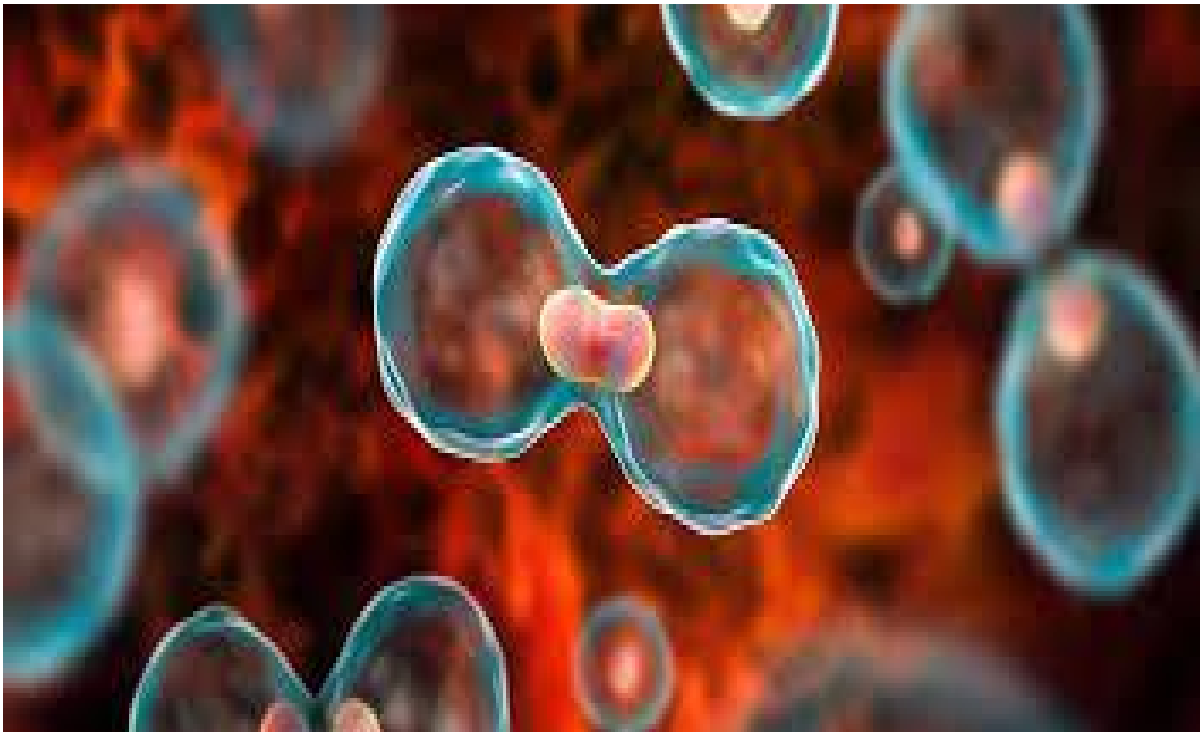
- Introduction
- Problem Statement And Goals
- About Dataset
- Data Preprocessing
- Exploratory Data Analysis
 - 1. Bivariate Analysis
 - 2. Univariate Analysis
- Classification
 - 1. Naive Bayes
 - 2. Decision Tree
 - 3. Support Vector Machine
 - 4. Random Forest
 - 5. K- Nearest Neighbor
- Conclusion

Introduction

Haberman's data set contains data from the study conducted in University of Chicago's Billings Hospital between 1958 to 1970 for the patients who underwent surgery of breast cancer.

I would like to explain the various data analysis operations I have done on this data set and how to conclude or predict survival status of patients who have undergone surgery.

Before building model or performing operation on data we should have good domain knowledge so that we can relate the data features and also can give accurate conclusion.



Problem Statement And Goals

By using this data my goal is to predict whether a patient will survive after 5 years or not based upon the patient's age, his/her operation_year and the number of positive lymph nodes

About Dataset

I would like to explain the features of the data set and how it affects other feature.

There are 4 attributes in this data set out of which 3 are features and 1 class label attribute as below. Also, there are 306 instances of data

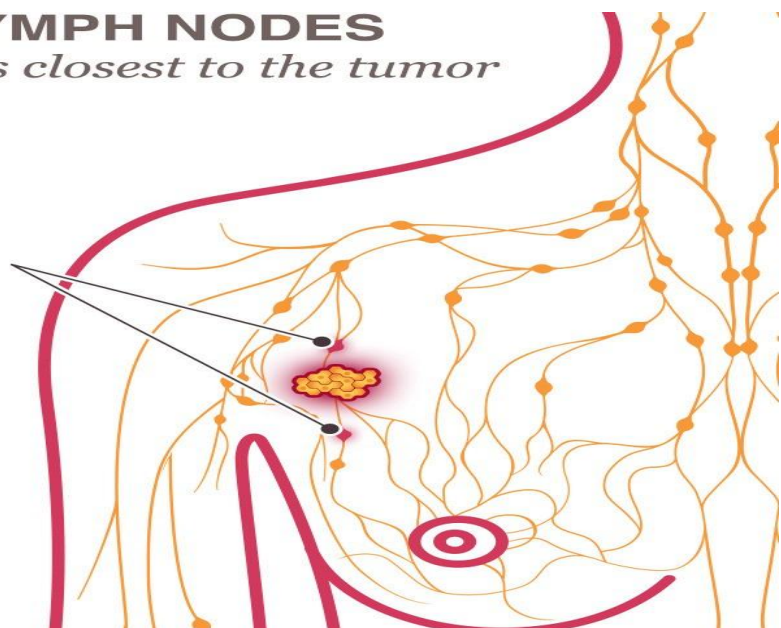
- Number of Axillary nodes(Lymph Nodes)
- Age
- Operation Year
- Survival Status

Lymph Node: Lymph nodes are small, bean-shaped organs that act as filters along the lymph fluid channels. As lymph fluid leaves the breast and eventually goes back into the bloodstream, the lymph nodes try to catch and trap cancer cells before they reach other parts of the body. Having cancer cells in the lymph nodes under your arm suggests an increased risk of the cancer spreading. In our data it is axillary nodes detected(0–52).

SENTINEL LYMPH NODES

the lymph nodes closest to the tumor

Sentinel
Lymph Nodes



Survival status (class attribute)

-- 1 = the patient survived 5 years or longer

-- 2 = the patient died within 5 year

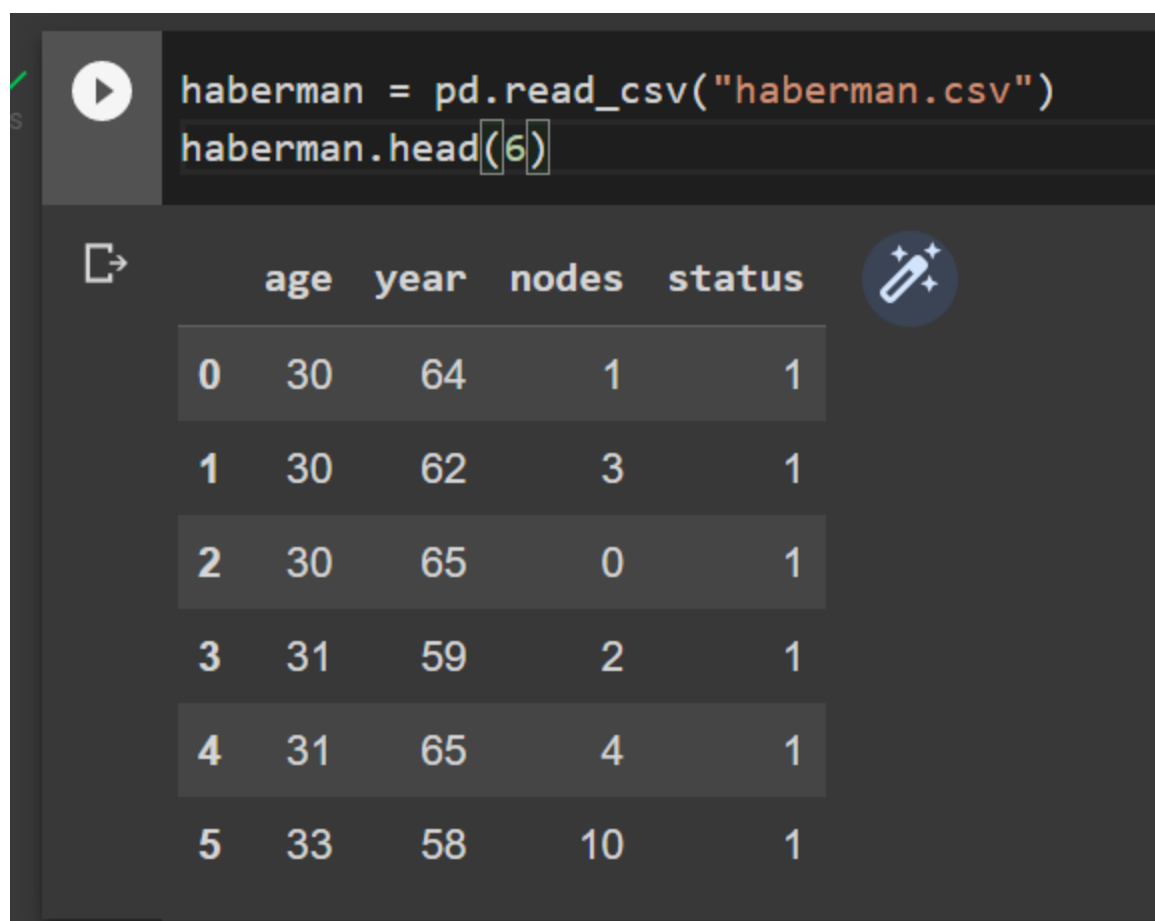
Age: It represent the age of patient at which they undergone surgery (age from 30 to 83)

Operation year: Year in which patient was undergone surgery(1958–1969)

Survival Status: It represents whether patients survive more than 5 years or less after undergoing surgery. Here if patients survived 5 years or more is represented as 1 and patients who survived less than 5 years is represented as 2.

Dataset Link :- <https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival>

This is how data looks like -



```
haberman = pd.read_csv("haberman.csv")
haberman.head(6)
```

	age	year	nodes	status
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1
5	33	58	10	1

Data Preprocessing

1.

```

haberman.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306 entries, 0 to 305
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype  
---  -
 0   age         306 non-null   int64  
 1   year        306 non-null   int64  
 2   nodes       306 non-null   int64  
 3   status      306 non-null   int64  
dtypes: int64(4)
memory usage: 9.7 KB

```

Observation:-we can clearly see that there are 4 columns age,year,nodes,status and there are no null and missing values in the whole dataset so there is no need of missing value treatment and we also saw that all the 4 columns contain integer data.

2.

```
haberman.describe()
```

	age	year	nodes	status
count	306.000000	306.000000	306.000000	306.000000
mean	52.457516	62.852941	4.026144	1.264706
std	10.803452	3.249405	7.189654	0.441899
min	30.000000	58.000000	0.000000	1.000000
25%	44.000000	60.000000	0.000000	1.000000
50%	52.000000	63.000000	1.000000	1.000000
75%	60.750000	65.750000	4.000000	2.000000
max	83.000000	69.000000	52.000000	2.000000

3.

```

haberman["status"].value_counts()

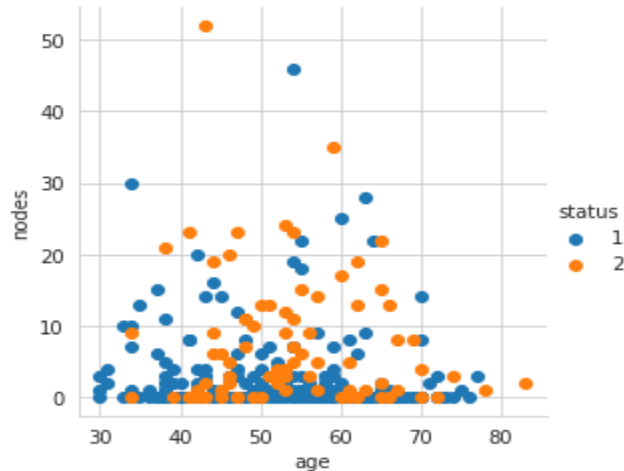
1    225
2     81
Name: status, dtype: int64

```

Observation:-It is a imbalanced dataset, in the data set there are 225 patients out of 306 survive more than five year after the surgery and 81 patient are survive less than five year after the surgery

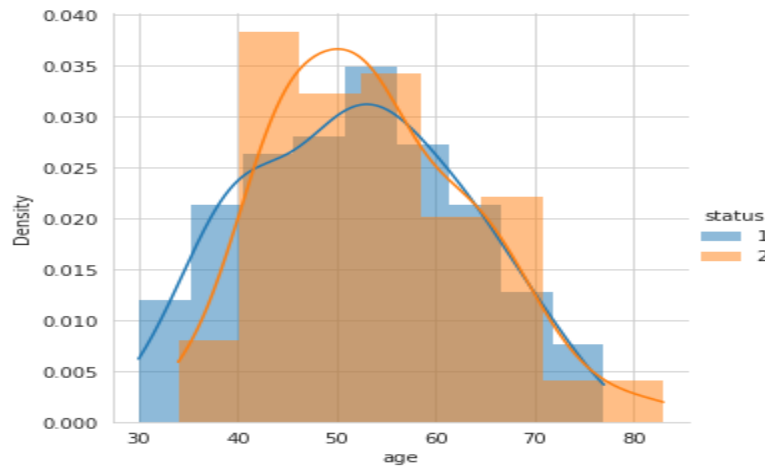
Exploratory Data Analysis

1.



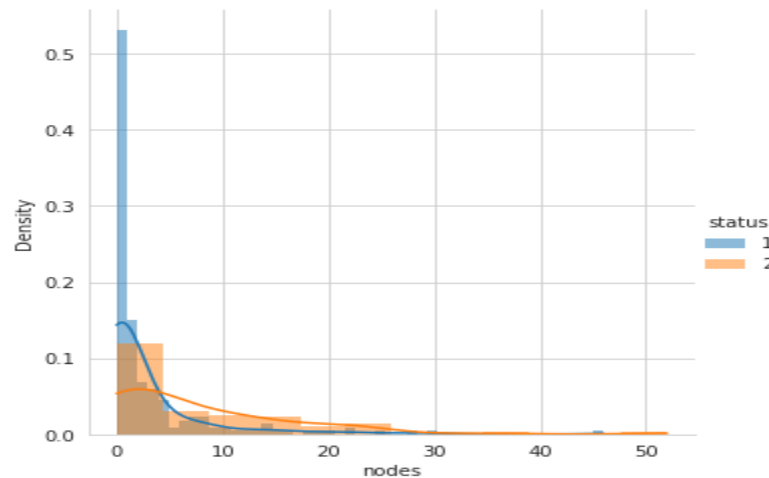
Observation:- From the above scatter plot we can make assumption that the person who have less nodes is higher chances to survive more than five year but still we are unable to distinguish 1 and 2.

2.



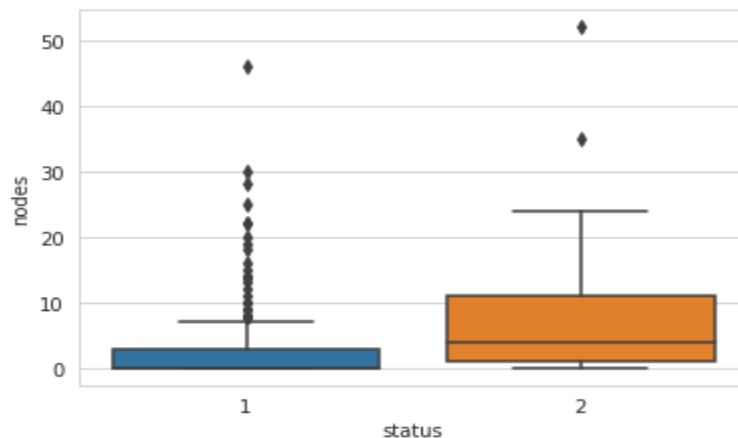
Observation:- From the above observation we can see that patient whose age are between 30 to 40 year the chances is little higher to survive more than five year and the patient whose age are between 40 to 56 year changes is little higher to died within five year but because most of the data are overlapped we are unable predict anything from this plot

3.



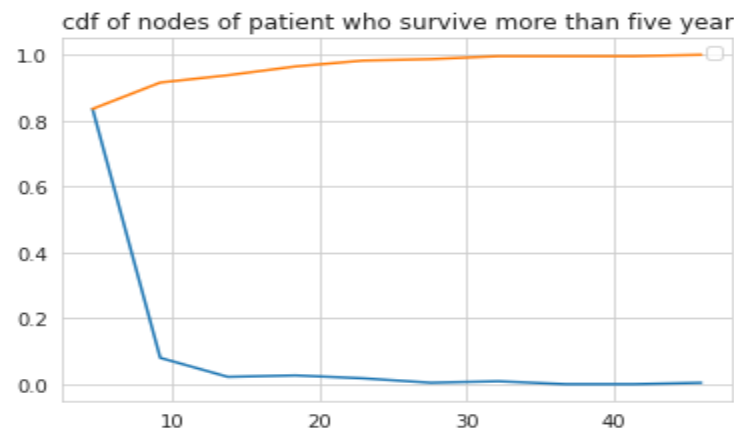
Observation:- From the above hist plot and pdf we can see that the patient with nodes<5 is higher chances to survive more than five year and as number of nodes increases the chances of survival decreases and chances of die within 5 year increases

4.



Observation:- From the above plot we can clearly see that the threshold for 1 is approximately 0 to 7. There are some outliers present in the data and 25th,50th percentile is the same for 1 which is 0 and 75th percentile is approximately 3. threshold for 2 is approximately 0 to 24, there are 2 outliers in 2 and 25th,50th,75th percentile are approximately 2,4 and 12 respectively. So from the graph we can see that if nodes<3 then chances is high to survive more than five year and as the number of nodes increases the chance to survive more than five year is reduces.

5.



Observation:- From the above plot we can clearly see that if less no of node the chances is high to survive (if nodes ≤ 3 chances of survival more than five year is approximate 82% as the number of nodes increases, chances of surviving more than five year decreasing very fast and after nodes ≥ 40 chances are very likely that person die within five year



Observation :- The feature nodes is more correlated with the class label as compared to other features means nodes is very important feature for classification whether a person survive after 5 years or not.

Classification

I applied different machine learning algorithms on this haberman cancer survival dataset and got different - different accuracy. The accuracy of different machine learning algorithms are as follow -

- Naive Bayes → 71.42%
- Decision Tree → 62.33%
- Support Vector Machine → 71.42%
- Random Forest → 71.42%
- K- Nearest Neighbor → 74.02%

Conclusion

We can diagnose the Cancer using Haberman's Data set by applying various data analysis techniques and using Machine learning algorithms.

Patient's age & year of operation is not a defining factor for survival. However, patients with age less than 35 are likely to survive.

Survival chances are more if less axillary nodes are present. Zero axillary nodes, however, do not guarantee survival.